Milestone 006 — A dramatic endpoint can still mislead

Tyler J.H.G.

Abstract

Does an apparent advantage survive fair checkpoint selection?

Research period: September 2–5, 2026
Historical anchors: #176, #177, #211

The unit model had a large advantage at the frozen 128-update endpoint. I kept investigating because an endpoint tells us where a model finished at a particular budget; it leaves open how it got there, whether the alternatives reached better checkpoints earlier, and what caused the gap.

Longer training, a matched regularization control, and a common checkpoint-selection policy changed the comparison. The original measurement stayed in the record, while the explanation became more specific.

1. Visual intuition

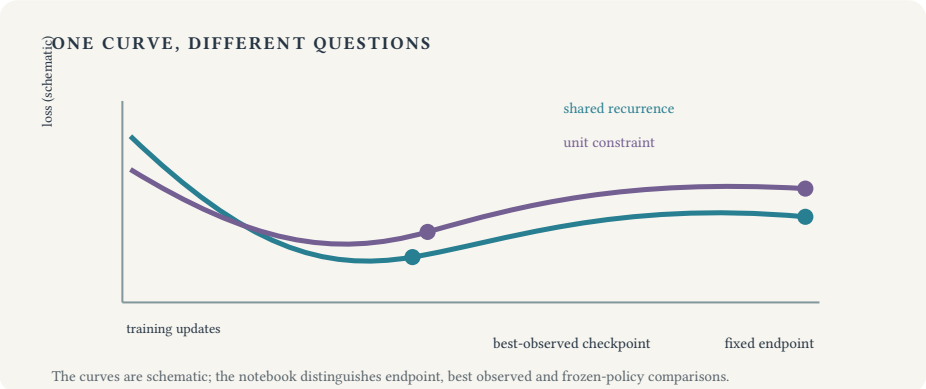

Figure 1: Schematic learning curves show why the best checkpoint and a fixed late checkpoint can rank models differently.

2. What changed

First, I extended the **training trajectory**. The shared and retained-radius models reached much better losses earlier; the unit model's advantage appeared later. The frozen endpoint had compared models at different places in their learning curves.

Published Sep 29, 2026

Second, I tested **regularization**. A Cartesian control matched the unit model's effects on vector norms and rank closely enough that the frozen decision rule returned

REGULARIZATION-EXPLAINS. Comparable regularization could reproduce the original endpoint effect, so that effect no longer required a geometry-specific explanation.

Third, I froze a common **checkpoint-selection policy** before comparing the models. Each architecture would be judged by the same rule, with the selection cost included in the budget.

Under that policy, selected validation NLL per byte was:

- shared recurrence: **3.3760**
- retained-radius recurrence: **3.3768**
- unit-hypersphere recurrence: **3.6205**

The unit model was about **0.2445 nat per byte worse** than shared recurrence (lower is better), with the simultaneous 95% interval entirely above zero. It also used about **3.44×** the mean training-plus-selection dense-operation budget.

The comparison had reversed. Once checkpoint selection followed the same rule, the unit model performed worse and cost substantially more.

> **Sticky note — checkpoint selection:** training produces many intermediate model states. A selection policy specifies which one is compared. Choosing that rule after inspecting the outcomes gives us another opportunity to tune the comparison around a favored result.

The historical data also needed a statistical correction. Some rows I had counted as separate observations came from the same underlying source article. Several measurements from one source can be useful, but they do not supply the independence of several different sources.

> **Sticky note — independent sampling unit:** the unit that contributes independent evidence to an analysis. Treating correlated measurements as independent makes the uncertainty look smaller than it is.

= Try a small example

Two toy loss sequences show how choosing the best checkpoint can change the comparison.

```
# Why selection policy matters: the best checkpoint and a fixed late
checkpoint can rank models differently.
shared = [3.9, 3.0, 3.4]
unit = [4.1, 3.5, 3.2]
print('best shared/unit:', min(shared), min(unit))
print('late shared/unit:', shared[-1], unit[-1])
```

Saved output:

```
best shared/unit: 3.0 3.2
late shared/unit: 3.4 3.2
```

3. Interpretation

A fixed endpoint, the best observed checkpoint, and a checkpoint selected by a predefined policy answer different questions. The #164 endpoint was a valid measurement of the first; the later comparison answered the third.

The matched control accounted for the motivating anomaly under the frozen rule, producing **REGULARIZATION-EXPLAINS**. Further geometric hypotheses need their own discriminating tests. The practical outcome here was a better comparison: trajectory analysis, common selection rules, matched regularization, and corrected source units.

4. Sources and chronology

- Milestone notebook 006 — preserved chronological source
- Research chronology
- Publication dispositions and Atlas-to-article map

The matching Atlas theme is **Checkpoint trajectories**. The earlier Visual Intuition Atlas notebook preserves the source visualization.

The next chronological article is **Milestone 007 — Derive before training**.